Binary Regression - The other way

STAT 245

Adjustment to our $\frac{n}{15}$ Rule

For all binary regression models

- We will definitely need a bigger dataset to estimate the probability of "success" *when success is very rare*
- Let s be the total number of successes in the dataset, and f the number of failures.
- Limiting sample size m is $min(s, f)$
- Number of coefficients we can estimate is about $\frac{m}{15}$

Thermal Preference

- Data from wearable sensors
- Can they predict whether people are cold?
- Define: success = to "Prefer Warmer"

Original Data

Like we're already used to

```
cold <- read.csv('https://sldr.netlify.app/data/cold.csv') |>
  na.omit()  |>
  glimpse()
```

```
## Rows: 2,974
## Columns: 10
## $ therm_pref   <chr> "Comfortable", "Comfortable", "Comfortable",
"Comfortable…
## $ location     <chr> "Indoor", "Indoor", "Indoor", "Indoor", "Outdoor",
"Indoo…
## $ sex          <chr> "Male", "Male", "Male", "Male", "Male", "Male",
"Male", "…
## $ exercise     <chr> "Low", "Low", "Low", "Low", "Low", "Low", "Low",
"Low", "…
## $ ambient_temp <chr> "Warm", "Warm", "Warm", "Warm", "Warm", "Warm",
"Warm", "…
## $ BMI_cat      <chr> "Moderate", "Moderate", "Moderate", "Moderate",
"Moderate
```

How many coefficients can we estimate?

```r
mosaic::tally(~therm_pref, data = cold)
```

```
## therm_pref
##   Comfortable Prefer Warmer
##          2381           593
```

Data Another Way

- Especially if we have categorical predictors, we can...
  - *group observations* and
  - tally up the **number of successes** and **number of observations** for all cases with *identical predictor variable values*.

Data "The Other Way"

Multiple trials per row

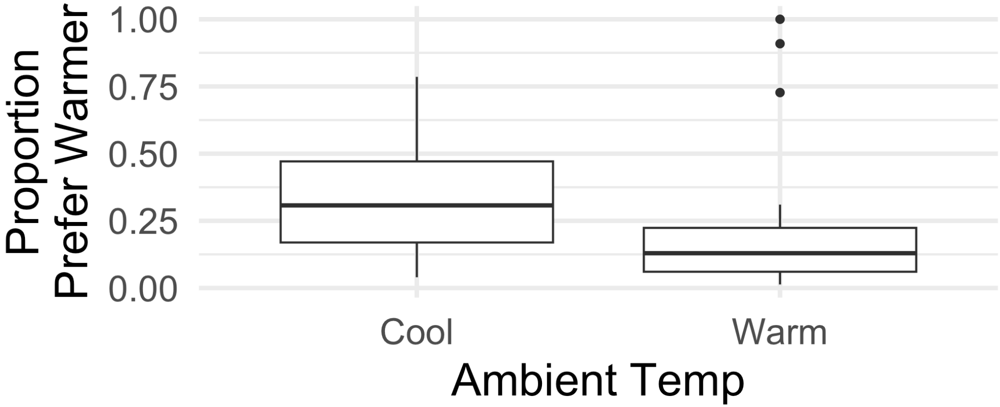
```
## Rows: 40
## Columns: 7
## $ location     <chr> "Indoor", "Indoor", "Indoor", "Indoor", "Indoor",
"Indoor…
## $ sex          <chr> "Female", "Female", "Female", "Female", "Female",
"Female…
## $ exercise     <chr> "High", "High", "Low", "Low", "Low", "Low",
"Moderate", "…
## $ ambient_temp <chr> "Cool", "Warm", "Cool", "Cool", "Warm", "Warm",
"Cool", "…
## $ BMI_cat      <chr> "Moderate", "Moderate", "Moderate", "Overweight",
"Modera…
## $ pref_warmer  <int> 35, 4, 20, 94, 17, 10, 10, 21, 12, 5, 7, 3, 128, 17,
14, …
## $ comfortable  <int> 161, 94, 96, 47, 98, 1, 70, 110, 157, 88, 69, 85, 261,
58…
```

Why???

- Maybe it came that way
- Easier to look at *proportion "success"* as a function of each predictor.

Easier Graphs

```
gf_boxplot((pref_warmer / (pref_warmer + comfortable)) ~
             ambient_temp,
           data = cold2) |>
  gf_labs(y = 'Proportion\nPrefer Warmer',
          x = 'Ambient Temp')
```



And linearity checking, too!

(If we had any quantitative predictors.)

```
gf_boxplot(logit(pref_warmer / (pref_warmer + comfortable)) ~
             quant_predictor,
           data = cold2) |>
  gf_labs(y = 'logit(Proportion\nPrefer Warmer)',
          x = 'Quant Predictor')
```

Binary Regression Setup

Multiple trials per row data

Use `cbind()` to group together the *number of successes* and *number of failures* to create the response variable.

```
cold_logit <-
  glmmTMB(cbind(pref_warmer, comfortable) ~
              location + sex + exercise +
              ambient_temp + BMI_cat,
            data = cold2,
            family  = binomial(link = 'logit'))
```

Logistic Regression - Results

`msummary()` - more concise than `summary()`

```
msummary(cold_logit)
```

```
##  Family: binomial  ( logit )
## Formula:
## cbind(pref_warmer, comfortable) ~ location + sex + exercise +
##     ambient_temp + BMI_cat
## Data: cold2
##
##      AIC      BIC   logLik deviance df.resid
##    378.8    392.3   -181.4    362.8       32
##
##
## Conditional model:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.76126    0.14161 -12.438  < 2e-16 ***
## locationOutdoor     0.74528    0.13055   5.709 1.14e-08 ***
## sexMale            -0.04033    0.10373  -0.389    0.697
## exerciseLow         0.71935    0.16257   4.425 9.65e-06 ***
## exerciseModerate   -0.11941    0.19244  -0.620    0.535
## ambient_tempWarm   -0.94035    0.10621  -8.854  < 2e-16 ***
## BMI_catOverweight   1.03279    0.13925   7.417 1.20e-13 ***
## BMI_catUnderweight  1.02872    0.17624   5.837 5.31e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The original way

One trial per row

```
old_cold_logit <-
  glmmTMB(factor(therm_pref) ~
        location + sex + exercise +
        ambient_temp + BMI_cat,
      data = cold,
      family  = binomial(link = 'logit'))
```

Summary, original way

One trial per row

```
msummary(old_cold_logit)
```

```
##  Family: binomial  ( logit )
## Formula:
## factor(therm_pref) ~ location + sex + exercise + ambient_temp +      BMI_cat
## Data: cold
##
##      AIC      BIC   logLik deviance df.resid
##   2699.4   2747.4  -1341.7   2683.4     2966
##
##
## Conditional model:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.76126    0.14161 -12.438  < 2e-16 ***
## locationOutdoor     0.74528    0.13055   5.709 1.14e-08 ***
## sexMale            -0.04033    0.10373  -0.389    0.697
## exerciseLow         0.71935    0.16257   4.425 9.65e-06 ***
## exerciseModerate   -0.11941    0.19244  -0.620    0.535
## ambient_tempWarm   -0.94035    0.10621  -8.854  < 2e-16 ***
## BMI_catOverweight   1.03279    0.13925   7.417 1.20e-13 ***
## BMI_catUnderweight  1.02872    0.17624   5.837 5.31e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compare coefficients (and SEs)

```
##                     Multi Trial Single Trial
## (Intercept)         -1.76125576  -1.76125576
## locationOutdoor      0.74527548   0.74527548
## sexMale             -0.04032832  -0.04032832
## exerciseLow          0.71934931   0.71934931
## exerciseModerate    -0.11940755  -0.11940755
## ambient_tempWarm    -0.94034675  -0.94034675
## BMI_catOverweight    1.03279267   1.03279267
## BMI_catUnderweight   1.02871876   1.02871876
```

One vs. Many Trials-per-row (don't do both!)

- Parameter estimates and SEs **identical**
- IC-based model selection *not identical*
  - Should we treat each observation of a success/failure as a draw from a binomial distribution with $n = 1$?
  - Should we treat each *set of trials with same predictor values* as a draw from a binomial distribution with $n \geq 1$?
  - Right answer depends on context, experimental design (beyond scope of our class?)

Pause: Odds Practice

The model equation for our model is:

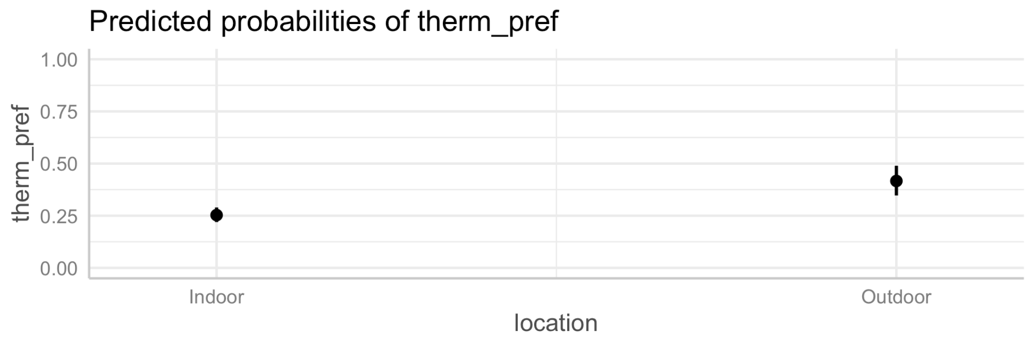
$$logit(p) = log\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 I_{outdoor} + \ldots$$

Where $I_{outdoor}$ is an indicator variable that is 1 when outside and 0 when inside, and our estimate of β_1 is $\hat{\beta}_1 = 0.745$ (from the model summary).

How do the odds of "Prefer warmer" change, when outside instead of inside?

Verification of Odds Interpretation

```
ggeffects::ggpredict(old_cold_logit,
                     terms = 'location') |>
  plot() |>
  gf_lims(y = c(0,1))
```



Notice: simpler to **just use predictions...**

plus odds when necessary

Model Assessment, Selection...

methods *same* regardless of data set-up :)

Other Links?

- may still use probit, cloglog if desired

Binary vs Count!

- Multi-trials-per-row binary data *can be mistaken for count data*
- For count data
  - **there is no "ceiling" (max possible count)**
- For binary data
  - **the number of trials is the "ceiling"**